Submitted to Recent Research Developments
in Molecular & Cellular Biology

Computational Discovery of Distinct RNA Elements with Functional Structures in Genomic Sequences

Shu-Yun Le and Jacob V. Maizel, Jr.

Laboratory of Experimental
and Computational Biology
NCI Center for Cancer Research
National Cancer Institute, NIH,
Bldg 469, Room 151, Frederick, Maryland 21702
Tel: 301-846-5532, Fax: 301-846-5598
e-mail: shuyun@ncifcrf.gov

Running Title: Functional Structured RNAs

ABSTRACT

Functional RNA elements (FRE) in post-transcriptional regulation of gene expression are often correlated with distinct RNA structures since FRE must fold into the specific conformations in which cell factors can recognize and interact with them. Recent computational studies indicate that the structures of FREs are both significantly more ordered and thermodynamically stable than anticipated at random. This is because of evolutionary constraints and intrinsic structural properties. Various computational tools for discovering well-ordered RNA structures and their structural motifs have been developed and a number of functional structured RNA elements have been determined. Here, we summarize recent efforts in the discovery of structured FRE within complex genomes by computation.

INTRODUCTION

Complete genomic sequence data are accumulating at an unprecedented pace. Nucleic acids sequences of pathogenic bacteria and human genomes provide the fundamental information useful for us to explore biological properties, such as regulation of gene expression. Although RNA is transcribed as a single-stranded, almost every RNA molecule has structure that includes various double helical, base paired regions formed by fold-back in the correct antiparallel orientation between complementary segments. In addition to Watson-Crick A:U and G:C base pairs, wobble G:U and other non-canonical base pairs also contribute to the structural constraints in the secondary and tertiary structure of RNA molecules.

Recent advances in studies of non-coding RNAs (ncRNAs) and RNA interference (RNAi) indicate that RNA is more than a messenger between genome and protein. The ncRNAs are involved in various regulatory mechanisms of gene expression at multiple levels [1-5]. Well documented instances include transcriptional mediation, RNA processing and modification, mRNA stability and localization, and translation of mRNA into protein [2, 6-10]. The functional structured RNAs (FSRs) that can perform the regulatory activity comprise transfer RNA, ribosomal RNAs, self-cleavage ribozymes [2], small microRNAs (miRNAs) [2, 3-5] and various RNA regulatory elements, such as iron-responsive element (IRE) in the non-coding region (NCR) of ferritin mRNAs [11], internal ribosome entry sequence in the 5' NCR [9] and cis-acting RNA elements involving in nuclear mRNA export, such as Rev response element (RRE) of HIV-1 and constitutive transport element (CTE) of Mason-Pfizer monkey virus [8]. The known biological functions of ncRNA continue to grow and newly discovered miRNA genes are one of the new classes of regulatory genes in animals. The ~ 22 nucleotides (nt) miRNAs can control gene expression by binding to complementary sites in the 3' NCR of target mRNAs [12]. It is interesting to note that miRNA precursors are of ~ 80 nt in length and form a conserved fold-back stem-loop structure across the divergent species in which the conserved ~ 22 nt miRNA sequences are within one arm containing at least 16 base-pairings [13]. Intriguingly, about a hundred distinct miRNA genes have been determined in *Caenorhabditis elegans* [36] and estimates for the number of miRNAs may range about $0.5 - 1$% of total protein-coding genes. It is conceivable that there are a large number of various FSRs in each genome. The FSR molecules are expected to be characterized by various structural motifs represented by specific combinations of base pairings and conserved nucleotides (nt) in the loop regions.

A complete understanding of a FSR requires a knowledge of its 3-D structure. The determination of its RNA 3-D structure is a limiting step in the study of RNA structure-function relationships because it is very difficult to crystallize and/or get nuclear magnetic resonance spectrum data for large RNA molecules. Currently, a reliable prediction of RNA secondary and tertiary structure from its primary sequence is mainly derived by phylogenetic comparisons with additional enzyme probing and the sensitivity of nucleotides to chemical modification [14-16]. The phylogenetic method has been demonstrated by successful predictions of RNA structures for tRNAs, 5S and 16S rRNAs, RNase P RNAs, small nuclear RNAs (snRNAs) and other RNAs, such as group I intron. Although dynamic programming and energy minimization methods [17-23] for predicting RNA structure are not as successful as phylogenetic comparative methods, they can be performed fast and automatically by computer. With improvements of the dynamic programming algorithm and parameters for the free energy of formation of RNA structural elements, $\sim 70\%$ of known base-pairs are predicted on average in optimized structures by the widespread MFOLD program [19-20]. The computed RNA secondary structures from MFOLD are often taken as working models that are further refined by multiple methods including experimental methods or phylogenetic comparisons. Moreover, computational methods for analysis and detection of FSRs have made a great progress recently. A number of tools [24-35] such as tRNAscan-SE, RNAMOT, Palingol, PatScan, Segfold, EDscan, SigED, RNAMotif, and ERPIN have been developed and have practical applications to the search for FSRs, such as tRNA genes, signal recognition particle and IRE. Some new computational procedures have been developed for identifying miRNAs encoded in worm and fly [36-37]. Here, we discuss recent efforts in the discovery of FSRs in genomic sequences by computation.

FSRs are Uniquely Folded

RNA structure comparison and analysis from numbers of laboratories show that some specific combinations of base pairings and some conserved loop sequences in stem-loops are more abundant in FSRs [38-41]. For example, analysis of a large number of ribosomal RNAs, such as 16S and 23S rRNAs, identified three classes of 4-nt terminal loops, or tetraloops of GNRA, UNCG and CUYG. In addition to rRNAs, GNRA tetraloops are also frequently found in self-splicing ribozymes and RNase P RNAs. The specific base-pairing and stacking implicated by non-canonical base-pairings G:A, A:G and R:R has also been found in loop E of eubacteria 5S rRNAs. The solution and crystal structures of *E. coli* 5S rRNA segments including loop E display a well-ordered structure characterized by the major groove narrowing and larger cross-strand distances in the central portion of loop E in which 4 out of 9 base-pairings are involved in non-canonical base-pairings [64]. It includes 2 G:A base-pairings as well as G:G and A:G. Also, the distinct base-pairing and stacking participated by non-canonical base-pairings and/or bulges are a common structural motif found in rRNAs, ribozymes and other various FSRs, such as IRE and HIV-1 regulatory elements RRE and TAR. On the other hand, phylogenetic conservation in FSRs is more impressive than that observed in the structural motifs of proteins. Statistical analysis indicates that there are about 15 invariant nts in a 76-nt tRNA molecule. Based on the observation of the distinct structural features it was suggested that FSRs possess well-ordered conformations that are both thermodynamically stable and uniquely folded [42].

To test the hypothesis, computational experiments were designed to explore the evolutionary constraints of the conformation folded in FSRs. Schultes et al. [43] computed three quantitative measures to estimate the stability and uniqueness of RNA secondary structures based on the mean length of stems and total number of base pairs in the predicted structure from RNAfold [18] and/or VIENNA [21]. The comparison of three scores computed from various FSRs and their randomly shuffled sequences indicates that the well-ordered conformations found in the most of FSRs are unlikely to arise from evolutionary modification only. Their results show that the well-ordered conformation of FSRs is expected to be rare in the conformation space formed from a population of the related random sequences.

It is evident that we must inspect the structure morphology in detail to evaluate the structural uniqueness more precisely. Recently, a novel algorithm (rna_match) for computing similarity between RNA structures was proposed [44]. In the structure comparison, each base-pair in helical duplexs and each nucleotide in single strands are examined and the maximal similarity score (MSS) between the two structures is computed. Using the quantitative measure MSS, the uniqueness of an arbitrary RNA structure can be estimated by evaluating the difference between the average MSS computed from a natural RNA and its related, randomly shuffled sequences and those MSS computed from random versus random sequences [45]. In the comparison, a standardized z-score $Stscr$ is introduced and defined as $Stscr = (RR - NR)/std$, where NR is the sample mean of computed MSS from the real RNA structure and a set of structures predicted from randomly shuffled sequences, RR and std are the sample mean and sample standard deviation of those MSS computed from the previous random structures versus the additional n random structures with the same composition and size as the natural RNA. Thus, the greater the $Stscr$, the statistically more unique is the well-ordered structure of the natural RNA.

In a computational test experiment on 100 tRNAs, the computed $Stscr$ were high and the $Stscr$

values averaged to 2.94 ± 1.02 indicating that the structural conformations of the natural tRNAs was significantly different from those of corresponding random structures, the uniqueness of the common cloverleaf structure was statistically significant. Also, the random tests for other FSRs including RNase P RNAs, TAR and RRE of HIV-1, IRES of HCV and ribozyme showed that the FSRs had well-ordered conformations that were unlikely to occur by chance. These tests strongly support the hypothesis that the well-ordered structures of FSRs are both thermodynamically stable and uniquely folded. It also indicates that the measurement of thermodynamic stability alone is not enough for us to characterize the structural features folded in the FSRs. FSRs consist of a well-ordered folding sequence (WFS).

Computational Strategy and Tactics in Finding FSRs

In addition to the development of efficient algorithms for predicting RNA high-ordered structure from the primary sequence, the another major goal of RNA structure computation is to discover potential FSRs in RNA sequences, and to correlate them with known experimental properties and to suggest candidate sites for further experimental study. Currently, there is no effective computational approach to detect FSRs that lack sequence or structure homology to one of the known FSRs. In general, computational prediction of potential FSRs in genomic sequences is further verified by experimental testing of expression levels, functional assay by deletion or mutagenesis and structural analysis. Currently, our computational strategy is often to delimit the potential FSR in a RNA sequence by searching for WFSs or unusual folding regions (UFRs). From the WFSs detected by a robust statistical inference we then explore the common structure features in homologous RNAs. Once the WFS is found to be both significantly stable and phylogenetically conserved it can be selected as a candidate for potentially FSR element. The homologous FSRs can be searched from sequence databases by pattern search tools based on both the primary sequence and the high-ordered structure of the experimentally verified FSR. Our tactics used in the procedure are summarized in Fig. 1.

Discovering WFSs in a Genomic Sequence

WFSs can be characterized by the thermodynamic stability and distinct conformation of the structure folded in local segments within a genomic sequence. Previously, WFSs were often searched by computer programs Sigstb and Segfold [33]. Sigstb and Segfold are used to explore a mRNA sequence by choosing successive fragments and comparing the computed free energy of the actual sequence to a number of randomly shuffled sequences of the same size and composition. The highly stable or unstable regions are statistically inferred and termed unusual folding regions (UFRs). It has reported that the detected UFRs in HIV-1, HIV-2, and other related viruses are

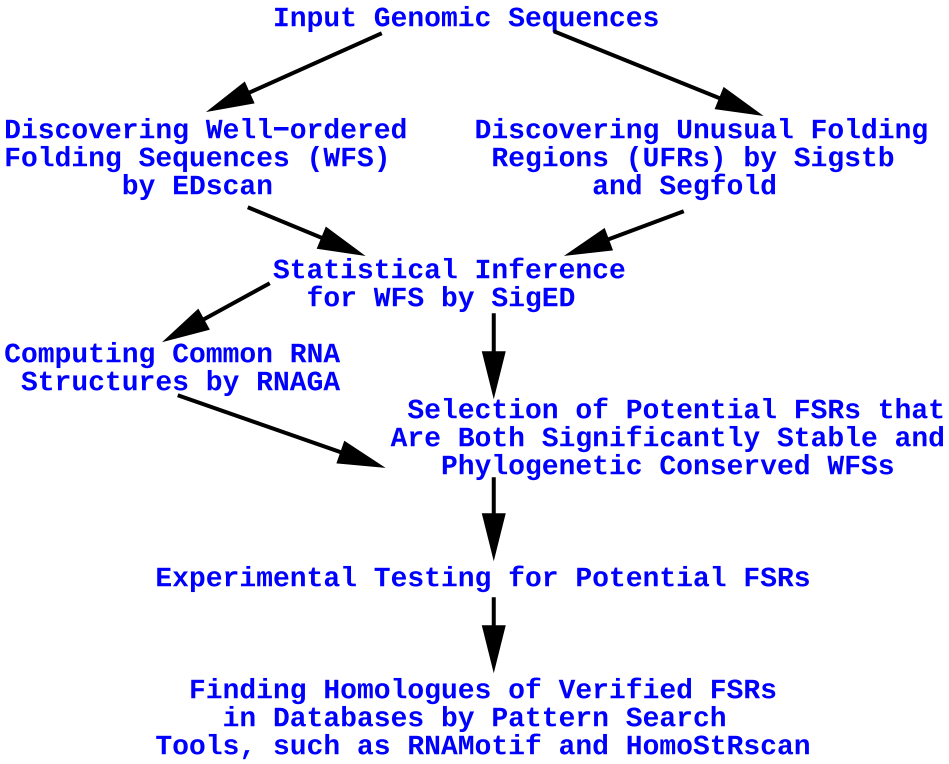
Input Genomic Sequences
Discovering Well-ordered Folding Sequences (WFS) by EDscan
Discovering Unusual Folding Regions (UFRs) by Sigstb and Segfold
Statistical Inference for WFS by SigED
Computing Common RNA Structures by RNAGA
Selection of Potential FSRs that Are Both Significantly Stable and Phylogenetic Conserved WFSs
Experimental Testing for Potential FSRs
Finding Homologues of Verified FSRs in Databases by Pattern Search Tools, such as RNAMotif and HomoStRscan

Figure 1: Procedure of discovering FSRs in genomic sequences. For details of the programs EDscan [34], Sigstb, Segfold [33], SigED [35], RNAGA [57], RNAMotif [28] and HomoStRscan [58] see previous publications.

coincident with the RRE and TAR[46-51].

In a recently developed computational tool EDscan [34], we used a quantitative measure E_{diff} to evaluate the quality of an arbitrary WFS. The measure $E_{diff}(S_i)$ of a given RNA segment (S_i)

is defined as the difference of free energies between the folded global minimal energy structure ($E(S_i)$) and its corresponding optimal restrained structure (ORS) in which all the previous base pairings in the global minimal structure are forbidden ($E_f(S_i)$). We have

$$\begin{aligned} E_{diff}(S_i) &= E_f(S_i) - E(S_i) \quad and \\ Zscr_e(S_i) &= \frac{E_{diff}(S_i) - E_{diff}(w)}{std(w)} \end{aligned}$$

where $E_{diff}(w)$ and $std(w)$ are the sample mean and standard deviation, respectively, of the E_{diff} scores computed by sliding a fixed-length window in steps of a few nt from 5' to 3' along a RNA sequence. It is clear that the score $Zscr_e(S_i)$ is a z-score, a standardized measure of $E_{diff}(S_i)$. We expect that the greater the $Zscr_e(S_i)$ of the segment S_i, the more well-ordered is the folded RNA fragment S_i.

EDscan utilizes a dynamic programming algorithm and Turner energy rules [18, 20] to compute $E_{diff}(S_i)$ and $Zscr_e(S_i)$ by scanning the RNA sequence. In searching for distinct WFSs in the sequence, we often take following steps. (i) $Zscr_e(S_i)$ is computed by sliding a window with a chosen size, for instance 80 nt in searching miRNAs, along the sequence. The potential interesting regions with high $Zscr_e(S_i)$ are chosen based on the statistical distribution of $Zscr_e(S_i)$ ($1 \leq i \leq N - W + 1$) computed by EDscan, where N and W are the length of the sequence and the sliding window. (ii) The precise locations of those potential targets in which the folded structure is highly well-ordered are inferred by an extended search in the regions determined from the step 1. In the extended search, the distributions of $Zscr_e$ in the selected regions are repeatedly computed by a set of windows whose size is systematically changed over a range of sizes (e.g. 60-100 for miRNAs). The maxima of $Zscr_e(S_i)$ are extracted to determine the optimized WFSs. (iii) The statistical significance of the computed WFSs is further tested by Monte Carlo simulation. For example, we may repeatedly compute the $Zscr_e(S_i)$ distribution in the randomly shuffled sequences using same procedure and parameters as used in the calculation of the natural sequence (see next section). The expected probability of a WFS detected in the natural sequence can be estimated from the random test.

Statistical Extremes of WFSs in the Sequence

To estimate the statistical extremes of WFSs in a long sequence, we need a good statistical model to describe the distribution of $Zscr_e$ in a large sample. Statistical analysis indicates that the $Zscr_e$ data show asymmetry with sample mean, $m = 0$, sample standard deviation, $std = 1.0$. The distribution of $Zscr_e$ is skewed toward the positive direction with a long tail and it is not well fitted by a normal distribution (see Fig. 2). To well estimate the statistical significance of $Zscr_e$ for a given RNA segment S_i we need to know what is the general behavior of $E_{diff}(RS_i)$ of a set of random

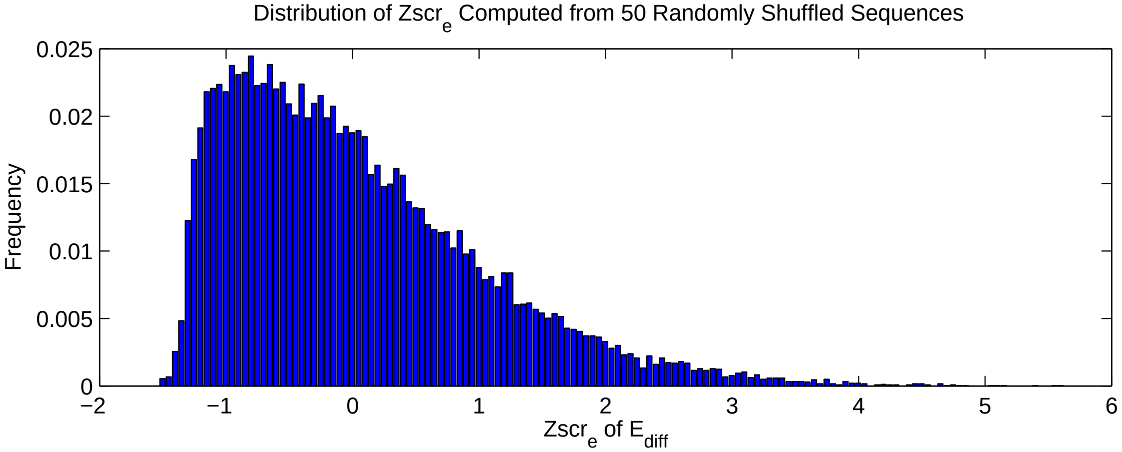

Figure 2: Empirical probability density functions of $Zscr_e$ scores computed from 50 random sequences of 2500-nt. The empirical bar functions are plotted with step size of $Zscr_e = 0.05$. $Zscr_e$ were computed by sliding a 80-nt window stepped with 5-nt each time along the randomly shuffled sequence.

sequences, $RS_{i,1}$, . . . ,$RS_{i,m}$, that are made by randomly shuffling the local segment S_i rather than the complete sequence. In a novel method SigED [35], a standard z-score, $SigZscr_e(S_i)$ was employed to evaluate the statistical significance of the energy difference $E_{diff}(S_i)$ computed from the segment S_i.

$$SigZscr_e(S_i) = \frac{E_{diff}(S_i) - E_{diff}(RS_i)}{std(RS_i)}$$

where $E_{diff}(RS_i)$ and $std(RS_i)$ are the sample mean and standard deviation of $E_{diff}(RS_{i,1})$, . . ., $E_{diff}(RS_{i,m})$. And $E_{diff}(RS_{i,1}), \ldots, E_{diff}(RS_{i,m})$ are the m values of energy difference computed from the m randomly shuffled sequences, $RS_{i,1}$, . . . ,$RS_{i,m}$. It is important to note that randomizations are done by shuffling so that the same base compositions and sizes as the natural fragment S_i are maintained. Similarly, we have

$$E_{diff}(RS_{i,j}) = E_f(RS_{i,j}) - E(RS_{i,j}) \quad (1 \leq j \leq m)$$

where $E(RS_{i,j})$ is the lowest free energy of the random fragment $RS_{i,j}$ and $E_f(RS_{i,j})$ is the minimal energy computed from the ORS of $RS_{i,j}$ in which all the previous base pairings formed in the lowest free energy structure are prohibited.

To facilitate statistical inference for distinct loops and base-pair stacking, the measure E_{diff} can be divided into two parts, $Estem_{diff}$ and $Eloop_{diff}$, to characterize the structural features of the base-pair stacking and loops, respectively. Where $Estem_{diff}$ is defined as the energy difference contributed by base-pair stacking only between the lowest free energy structure and its corresponding ORS. $Eloop_{diff}$ is defined as the energy difference contributed by loops only between the two structures as mentioned above. Furthermore, we can define the other two standardized z-scores, $SigStem_e(S_i)$ and $SigLoop_e(S_i)$ for the given segment S_i [35]. The two z-scores, $SigStem_e(S_i)$ and $SigLoop_e(S_i)$ are helpful in discovering distinct loops and significantly unstable folding regions. The method SigED is used to infer statistical extremes of WFSs by computing $SigZscr_e$, $SigStem_e$ and $SigLoop_e$ with scanning successive segments along a nucleotide sequence.

Prediction of Common RNA Secondary Structures

A number of computational methods [52-56] for predicting common RNA secondary structure for a set of related RNA sequences have been proposed. Most of these methods start with a set of predicted RNA structures computed from their related RNA sequences by a thermodynamic dynamic programming algorithm and their multiple sequence alignment. The predicted base-pairings are gradually refined by the analysis of sequence covariation or mutual information so that a common RNA structure for the set of RNAs is emerged. The recently developed program RNAGA [57] is different from other approaches. The method RNAGA employs a genetic algorithm (GA) to search for a common secondary structure without the need for pre-aligned homologous RNA sequences. One of the remarkable features of RNAGA is that RNA secondary structures are automatically optimized by not only the free energy of the formation of the structure but also the structural similarity among homologous sequences [40]. The program is a three-step procedure. In the first stage, a GA is used to generate a population of RNA secondary structures that satisfy certain conditions of thermodynamic stability. In this step, the free energy of a folded structure is taken as a fitness criterion. Secondly, the structural similarity between any two structures within the population of RNA secondary structures is computed. With the quantitative measure of structural similarity as the fitness criterion, a GA is then used to improve the structural similarity among homologous RNAs for the structures in the population of a sequence. Finally, those structures that satisfy certain conditions of thermodynamic stability and structural conservation are selected as predicted common structures for a set of homologous RNAs. As a result, RNAGA solves the alignment problem of multiple sequences and the folding problem of common RNA structures simultaneously. The program also ranks the predicted common structures based on the structural similarity score in descending order. In a test including a set of 20 tRNA sequences, 25 5S rRNAs, 7 HIV-1 RREs and 10 RRE of HIV-2 and SIV, fairly convincing common secondary structures were obtained by RNAGA in the top 10 ranked solutions [57].

In the method, a secondary structure is considered as an individual in the population. A structure is encoded as a set of stems, such as $T = \{s_1, s_2, \cdots, s_n\}$. A random structure in the sequence is produced by randomly choosing a stem s_i from the stem list consisting of all possible stems occurred in a sequence. In the structure construction in the first step of the approach, a stem can be added to the structure only if the addition of a stem increases the structure stability, otherwise the addition is determined by the Boltzmann rule. The process is repeated again and again until no more such stem s_i can be added from the stem list. In the optimization, RNAGA operates on a population of tentative solutions by crossover and mutation operators. Thus, an offspring of the two parental structures is constructed by crossover and/or mutation on the parental structures.

Database Search for RNA Structural Motifs

Over the last decade the computational search methods for distinct RNA structural motifs have made great progress. A number of database search tools have been developed and have practical applications to the search for known FSRs and their homologues [24-32, 58]. In general, these pattern search tools can be divided into two groups. Tools in the first group are designed to search for a specific FSR, such as tRNAs. Among them, the method tRNAscan-SE is very efficient and successful in finding tRNA genes in complete genomes [32]. The methods in the second group are designed and optimized to find general RNA structural motifs. Most of these algorithms provide a descriptor that can describe the RNA structural elements of known FSRs and a pattern search algorithm to match and score the patterns found in the genomic sequence. The recently developed algorithm, RNAMotif [28] is more powerful and efficient than others. A significant improvement in RNAMotif is that its descriptor can specify any type of base-base interaction and RNA structural element. Also RNAMotif provides a user controlled scoring system that can be used to add capabilities in the pattern matching.

The main shortcoming of pattern-based search tools is a general inability to incorporate information of sequence and structural feature in detail. Holbrook and colleagues [63] have proposed a general computational approach to identify FSRs in genomic sequences using neural network simulations. We recently developed a novel algorithm, HomoStRscan, of searching for homologous FSRs by scanning a genomic sequence [58]. HomoStRscan differs from other currently used approaches in considering each base and base pair in the query RNA. Among them, any type of base-base interaction is allowable. The algorithm finds the most similar structure to match the query structure in an arbitrary segment in the target sequence. The size of the arbitrary segment ranges near the length of the query RNA, and can be flexibly controlled by users. Simultaneously, the MSS between the query RNA and each computed matching structure from the target sequence is calculated. The homologous RNAs are then predicted by robust statistical inference from the MSS distribution computed by moving a window along the target sequence. Thus, HomoStRscan can be used to search in the genomic sequence for any RNA motif corresponding to an established secondary structure. Computational test experiments for several complete bacterial genomes proved to be very effective in finding ncRNAs, such as tRNAs and 5S rRNAs [58].

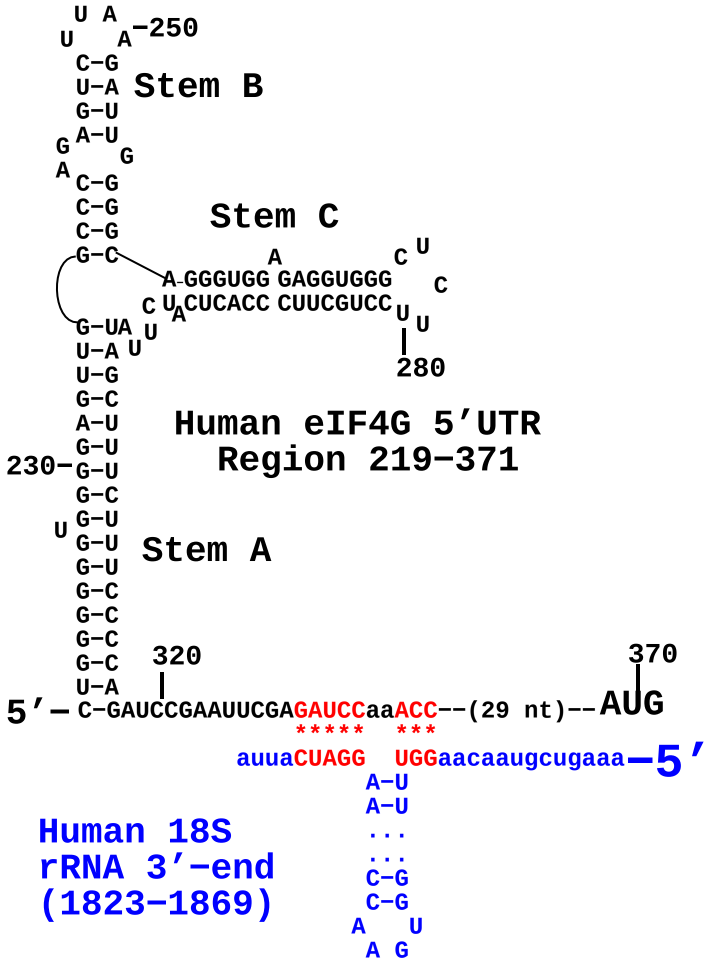

Figure 3: Y-shaped stem-loop structure computed in the 3’ portion of the 5’UTR of human eIF4G mRNA. The Y-shaped motif is denoted by Stem A, B and C. The short 18S rRNA-complementary sequence is labeled by the character *.

Discovering Functional RNA Elements

Functional Structured Elements in the 5'UTRs

Experimental studies revealed that a special region called the internal ribosome entry segment (IRES) allowed the translational machinery to skip over upstream AUGs [9]. The IRES elements detected in cellular mRNAs are quite divergent and their sizes range to initiate translation at the correct codon from ~ 100 nt in human immunoglobulin heavy chain binding protein mRNA to ~ 630 nt in the 5'UTR of B chain of human platelet-derived growth factor. The predicted common structural core in these cellular IRES elements shows a distinct Y-shaped stem-loop structure (a 3-way junction) [59-61]. While it is still true that most mRNAs initiate translation from their first AUG there are a growing number of interesting cases where internal initiation plays a role in regulation of expression at a post-transcriptional level. A statistical analysis of the upstream AUG (uAUG) in a database of 5'UTRs, UTRdb [62], indicated that $\sim 56\%$ of human mRNAs have no uAUG in the 5'UTR and 901 out of 6669 ($\sim 14\%$) human mRNAs have three and more uAUG codons. We found that a number of mRNAs of oncoproteins, growth factors, transcription factors, signal transduction genes and immune or inflammation mediators have a long, GC-rich and structured 5' UTR with multiple uAUG. Using the integrated approach shown in Fig. 1, we found a common Y-shaped stem-loop followed by a short, 18S rRNA-complementary sequence immediate to the initiator in a number of cellular IRESs and other long 5'UTRs of high G+C content and multiple uAUGs (see Fig. 3, Tables 1 and 2). This common structural motif is suggested to be associated with the important biological role of reported cellular IRESs.

Fold-back Stem-loops of the Reported miRNA Precursors are Coincident with Statistically Significant WFSs

The genome of *C. elegans* is organized into six chromosomes with total size of about 100 million nts. We computed the $Zscr_e$ distribution by EDscan [34] by scanning the fixed-length window of 80-nt with a step of 5 nt along each chromosome sequence. We found some interesting noncoding regions in which the computed WFS elements with very high $Zscr_e$ were clustered. As shown in Fig. 4 we detect those WFSs that are coincident with the well known miRNAs, mir-35, mir-37, mir-38, mir-39, and mir-40.

Using the profile of computed $Zscr_e$ in the genomic sequences we can further refine the analysis of WFS by SigED [35]. The best $SigZscr_e$ scores of WFSs that are associated miRNAs are summarized in Table 3. Our results indicate that most $SigZscr_e$ scores are greater than 3.5 and their expected random probability is less than 0.0002. It shows that the fold-back stem-loops folded by the precursors of known miRNAs in *C. elegans* genome are closely associated with the statistically significant WFSs. With the additional information, such as miRNA phylogenetic conservation, EDscan and SigED can be used to search for ncRNAs in genomes.

5'UTR	Size (nt)	No. of uAUG	Y-shaped Structural Motif A	 B	 C	 D	Complementary Sequence
Human 5'UTR having Cellular IRESs							
AML1	1580	15	1469-1480 //1546-1558	1482-1520	1521-1545	N	CUUGUUGUG(0 nt)AUG
BiP	221	0	129-142 //183-194	144-160	161-179	N	ACuGGCU(6 nt)AUG
C-myc	513	0	228-249 //340-368	252-311 (CUG is an alternative initiator for C-myc)	313-336	Y	UGcUUAGAC(1 nt)CUG
eIF4G	368	4	219-236 //301-317	237-259	260-296	N	GAUCCaaACC(29 nt)AUG
FGF-2	466	3	204-221 //263-276	222-241 (CUG is an alternative initiator for FGF-2)	242-259	N	GCGGCU(5 nt)CUG
			355-369 //420-437	371-391	392-418	N	GGgGAUCCcgGCC(16 nt)AUG
PDGF2 /c-sis	1022	3	941-944 //992-995	946-969	970-990	N	GCCcggaguCGGC(0 nt)AUG
VEGF	1038	1	845-855 //978-987	858-907	909-977	Y	GGCCUCC(6 nt)AUG
Human Cellular 5'UTRs							
abl (M14753)	340	6	225-236 //309-321	237-285	289-307	N	GGU--ACC-UAUUAUuACUUU -(0 nt)AUG
c-abl (M14752)	364	0	282-294 //339-355	295-320	321-338	N	UGGcGcAA-A(0 nt)AUG
bcr (X02596)	488	2	373-383 //445-455	385-404	405-437	N	GGCGG--CGC(9 nt)CGGC -(6 nt)AUG
c-erb (Y00479)	333	3	118-135 //232-249	138-191	192-231	Y	GGcAUCC(9 nt)UUGaa -GUGA(0 nt)AUG
c-erbA-1 (X55005)	466	4	248-266 //365-382	269-323	324-364	Y	GGcAUCC(9 nt)UUGaa -GUGA(0 nt)AUG
IL-15 (X91233)	316	10	221-233 //279-290	234-244	245-277	N	UAAgGAUUUACC--GU ----GGCUUU(5 nt)AUG
Int-2 (X14445)	491	3	396-406 //445-456	410-425	426-442	Y	GAUGCC(3 nt)AUG
mas (M13150)	267	3	117-131 //235-249	134-194	197-234	N	CCaACCU-GaGGCcU -(4 nt)AUG
mos (J00119)	479	1	369-381 //461-473	383-437	439-452	N	AUcAUC(0 nt)AUG

Table 1: Y-shaped structural motif and a short complementary sequence to the 3' end of human 18S rRNA sequence found in the cellular IRESs and some large cellular 5'UTRs that contain multiple upstream AUG. The folding regions of stems A, B, and C in the Y-shaped motif (see Fig. 3) are listed in the columns 4, 5 and 6. An additional stem-loop D between the Y-shaped motif and 18S rRNA-complementary sequence is denoted by letters Y (Yes) and N (No) in the seventh column. The 18S rRNA-complementary sequences are represented by capital letters in the last column. All of complementary sequences observed in human 18S rRNA are located at the upstream and/or downstream single-stranded regions (1823-1838 and 1861-1869) of the folded hairpin structure (1839-1860) in the 3'-end as shown in Fig. 3.

5’UTR	Size (nt)	No. of uAUG	Y-shaped Structural Motif A	B	C	D	Complementary Sequence
Mouse and Other Cellular 5’UTRs							
abl-2 (U13835)	144	3	55-68 //115-125	70-93	95-114	N	UCCaGCCUcCGAC(0 nt)AUG
abl-3 (X07539)	219	0	91-101 //155-166	102-136	137-153	N	UAA-GGUCCuugugaGCC -acgUUGUGGU(25 nt)AUG
abl-4 (X07541)	1168	11	1059-1068 //1134-1145	1069-1116	1119-1133	N	CACCUaUUAUuGCUUU -(0 nt)AUG
Rat BiP (M14866)	206	0	114-126 //158-169	128-141	142-157	N	CCGCUgagcgACuGACU -(19 nt)AUG
Hamster BiP (M17169)	150	0	64-73 //114-125	74-88	89-110	N	GGCCCACagcGCcGGC -(3 nt)AUG
int-2 (Y00848)	864	3	776-788 //821-833	789-805	806-817	Y	GAUGCC(3 nt)AUG
Rat FGF-2 (M22427)	532	0	310-321 //363-375	323-337 (CUG is an alternative initiator for FGF-2)	338-362	N	GUCCgGCU(8 nt)CUG
			438-450 //497-510	452-469	470-493	N	GUCCcgggGCC--- ---GCGG(7 nt)AUG
Rat C-myc (Y00396)	413	0	71-92 //194-222	94-139 (CUG is an alternative initiator for C-myc)	143-188	Y	UUAUU-UGA(3 nt)CUG
mas (U96273)	341	5	87-105 //193-211	107-145	146-189	N	CCACCg(0 nt)AUG
mos (X12449, Monkey)	479	1	369-381 //461-473	383-437	439-452	N	AUcAUC(0 nt)AUG
Rat mos (X52952)	482	2	369-384 //464-477	386-440	441-458	N	UAAUc(0 nt)AUG
Chicken mos (M19412)	487	4	377-389 //469-481	391-445	448-463	N	AUcAUC(0 nt)AUG
Xenopus mos (X13311)	483	2	373-385 //461-477	387-441	444-458	N	AUcAUC(0 nt)AUG
VEGF (U41383)	1014	1	818-829 //952-962	832-880	883-951	Y	AcGGcCU-CC(6 nt)AUG
Bovine VEGF (M32976)	533	0	352-362 //485-494	365-413	416-484	Y	caGGcCU-CC(6 nt)AUG
Yeast eIF4G (L16923)	528	11	438-450 //493-507	453-469 (Yeast TIF4631)	470-491	N	ACCUaUUAC(4 nt)AUG
Yeast eIF4G (L16924)	528	4	437-453 //490-502	455-467 (Yeast TIF4632)	468-486	N	AAUaGAUCaaUUGU-Ag- GcACU(0 nt)AUG

Table 2: Y-shaped structural motif and a short complementary sequence to the 3’ end of human 18S rRNA sequence found in the cellular IRESs and some large cellular 5’UTRs that contain multiple upstream AUG. The folding regions of stems A, B, and C in the Y-shaped motif (see Fig. 3) are listed in the columns 4, 5 and 6. An additional stem-loop D between the Y-shaped motif and 18S rRNA-complementary sequence is denoted by letters Y (Yes) and N (No) in the seventh column. The 18S rRNA-complementary sequences are represented by capital letters in the last column. All of complementary sequences observed in human 18S rRNA are located at the upstream and/or downstream single-stranded regions (1823-1838 and 1861-1869) of the folded hairpin structure (1839-1860) in the 3’-end as shown in Fig. 3.

Table 2. The reported miRNAs in *C.elegans* genome and their corresponding WFS determined by EDscan and SigED.

Gene	Corresponding Well-ordered Folding Sequences (WFS)	SigZscr
lin-4	UUCCCUGAGACCUCAAGUGUGAGUGUACUAUUGAUGCUUCACACCUGGGCUCUCC	3.34
let-7	UGUGGAUCCGGUGAGGUAGUAGGUUGUAUAGUUUGGAAUAUUACCACCGGUGAACUAUGCAAUUUUCUACCUUACCGGAG	8.35
mir-1	GUGACCGUACCGAGCUGCAUACUUCCUUACAUGCCCAUACUAUAUCAUAAAUGGAUAUGGAAUGUAAAGAAGUAUGUAGAACGGGGUGGU	7.01
mir-2	CAUCAAAGCGGUGGUUGAUGUGUUGCAAAUUAUGACUUUCAUAUCACAGCCAGCUUUGAUG	3.85
mir-34	AGAGGCAGUGUGGUUAGCUGGUUGCAUAUUUCCUUGACAACGGCUACCUUCACUGCCACCCCGAACAUGUCGUCC	5.11
mir-35	UCAGAUCGAGCCAUUGCUGGUUUCUUCCACAGUGGUACUUUCCAUUAGAACUAUCACCGGGUGGAAACUAGCAGUGGCUCGAUCUUUUCC	9.49
mir-36	GUCGGGGAACCGCGCCAAUUUUCGCUUCAGUGCUAGACCAUCCAAAGUGUCUAUCACCGGGUGAAAAUUCGCAUGGGUCCCCGAC	7.36
mir-37	CCCUUGGACCAGUGUGGGUGUCCGUUGCGGUGCUACAUUCUCUAAUCUGUAUCACCGGGUGAACACUUGCAGUGGUCCUC	5.55
mir-38	AGGUCCUGUUCCGGUUUUUUCCGUGGUGAUAACGCAUCCAAAAGUCUCUAUCACCGGGAGAAAAACUGGAGUAGGACCUG	10.01
mir-39	GAGAGCCCAGCUGAUUUCGUCUUGGUAAUAAGCUCGUCAUUGAGAUUAUCACCGGGUGUAAAUCAGCUUGGCUCUGGUGU	6.97
mir-40	CCGCACCUCAGUGGAUGUAUGCCAUGAUGAUAAGAUAUCAGAAAUCCUAUCACCGGGUGUACAUCAGCUAAGGUGCGGGU	6.85
mir-41	UCCCAGAGACCUUGGUGGUUUUUCUCUGCAGUGAUAGAUACUUCUAACAACUCGCUAUCACCGGGUGAAAAAUCACCUAGGUCUGGAGCC	3.45
mir-42	GGACCUUUGUGGGUGUUUGCUUUUUCGGUGAAGUUGUCUUCCGUAGCUUCUUCUUCACCGGGUUAACAUCUACAGAGGUCCAAAAAGGGG	7.80
mir-43	GCCCGUGACAUCAAGAAACUAGUGAUUAUGCCAAACCACAGGGACAUAUCACAGUUUACUUGCUGUCGCGGGCGG	10.39
mir-44	GGCCAAUCUGGAUGUGCUCGUUGGUCAUAGACGUCAACACGAACUGUUCAUAUGACUAGAGACACAUUCAGCUUGGCCUG	6.76
mir-45	GUGCCACGCUGGAUGUGCUCGUUAGUCAUAAUAUCCUCCACAAAGCAAGGACUAUGACUAGAGACACAUUCAGCUUGGCG	9.80
mir-46	GCUGAAGAGAGCCGUCUAUUGACAGUUCAAGACCACGAGUCGUUGUGUGCUGUCAUGGAGUCGCUCUCUUCAGAU	5.55
mir-47	AAACUGAAGAGAGCAGUCUAUUGACAGUCGGUUACUCGAAAUCUUUACUGUCAUGGAGGCGCUCUCUUCAGAUGA	7.88
mir-48	aactctgggaatgcgagctaggctggtggatgtgagataccgttcaatTCGCATCTACTGAGCCTACCTCAagttcccgggagtt(antisense)	8.17
mir-49	AAAAGACCACCGUCCGCAGUUUGUUGUGAUGUGCUCCAAGCAAUCAUGAGUCUGAAGCACCACGAGAAGCUGCAGAUGGAGGUUC	3.86
mir-50	UGCCCGCCGGCCGCUGAUAUGUCUGGUAUUCUUGGGUUUGAACUUCCAGCGUUGAACCCGCAUAUUAGACGUAUCGACGGCCGGCGGGGC	10.32
mir-51	CGUCUACCCGUAGCUCCUAUCCAUGUUACUGGUCAAAAAGUGAACAUGGAAGCAGGUACA	3.84
mir-52	UCCAACUCUAACAGUCCACCCGUACAUAUGUUUCCGUGCUUGACAGCGAAGCUCAAUCACGUUACAAUGAAAGGGUAGCCGGUUAUUGAAGUUGG	3.24
mir-53	ACCCGUACAUUUGUUUCCGUGCUUGACUUCAAAGCUCAAUCACGGCACAAUAUAUGGGUC	3.55
mir-54	CGCUCUGACUAGGAUAUGAGACGACGAGAACAUUGCUUUUUUAAAAGACUUGUACCCGUAAUCUUCAUAAUCCGAGUCAGGGCUAGCUGA	5.49
mir-55	GGGACUCGGCAGAAACCUAUCGGUUAUACUUUUUGGAUAUGCUAUACCCGUAUAAGUUUCUGCUGAGCCCCUUAU	7.95
mir-56	CUGUUCUUGGCGGAUCCAUUUUGGGUUGUACCUCAUCCUAAAUUUGACGGUACCCGUAAUGUUUCCGCUGAGAACCGACU	7.51
mir-57	CUACCCUGUAGAUCGAGCUGUGUGUUUGAAACAAUCAUACACGAGCUAGACUACAAGGUGCACGAACAAACCGAA	4.39
mir-58	CAUAUCCAUUGCCCUACUCUUCGCAUCUCAUCACUUCGUCCAAUACCAUAGGGAUGAGAUCGUUCAGUACGGCAAUGGAC	5.52
mir-59	UAUGACAUCGUCCUGAAAACGAAACGGAACAAAAGUUCAAGAUAUUGAUUUCGAAUCGUUUAUCAGGAUGAUGUG	5.63
mir-60	UCUUGAACUGGAAGAGUGCCAUAAAAUCAUGACAAAGUACGUGAUAUUAUGCACAUUUUCUAGUUCAAGACUUGA	10.43
mir-61	UAUCGCUGAACCUCGAGAUGGGUUACGGGGCUUAGUCCUUCCUCCGUAUGGCAAUGACUAGAACCGUUACUCAUCUCGAGGUUUCGGUGA	8.11
mir-62	GGUGAGUUAGAUCUCAUAUCCUUCCGCAAAAUGGAAAUGAUAUGUAAUCUAGCUUACAGG	5.18
mir-63	GACACAAUUUCUAACUCGUCGGUAGUCAUCGUUCUAGCUGAAAAGGACACUAUGACACUGAAGCGAGUUGGAAAUAGUGGUUCUA	8.24
mir-64	CGCCGAAUAUGACACUGAAGCGUUACCGAACCGUUUUCCCACACCUGGAUUCGGUGCAACGAUCAGUGGCAUGCUCGGCU	5.70
mir-65	AUGGAGCCUUCGCCGAUUAUGACACUGAAGCGUAACCGAACACCAUAUUUUGAGAUUCUG..(25 nt)..GUUGGCUCCAUUAAA	4.09
mir-66	CCACAAAAAUGCCAUACAUGACACUGAUUAGGGAUGUGAUGAAUGUUAAGAUCCCGAUCA..(20 nt)..AUGGCGUAUGUGGUU	7.20
mir-67	GUCGAUCCGCUCAUUCUGCCGGUUGUUAUGCUAUUAUCAGAUUAAGCAUCACAACCUCCUAGAAAGAGUAGAUCGAUUUU	9.96
mir-68	UUUUGAAAUUCAUUUUUCUGAAUUUCACACUUUCAGUUAGUUGAUAUUAACGUUUGUAAAUAGGAUGGUAUAUUCGAAGACUCAAAAGUGUAGAC	3.06
mir-69	UUAAUUUAAUUUUUUUUUAAUUUUUAACGGGGUUAUUCAAGUAAUAUCGAAAAUUAAAAAGUGUAGACAU	3.33
mir-70	UCAAAAUAAA..(25 nt)..CGACGAAUAACACUUAUGAAGAAAUGUAAUACGUCGUUGGUGUUUCCAUAGUUUGAAUUGUUUAU	4.15
mir-71	CUGCUCUGAACGAUGAAAGACAUGGGUAGUGAGACGUCGGAGCCUCGUCGUAUCACUAUUCUGUUUUUCGCCGUCGGGAU	4.57
mir-72	GGUCCCGUCAGAGCUAGGCAAGAUGUUGGCAUAGCUGAAUGAUCGCUAUAACAACUAUCAGCUUCGCCACAUUCUGCCACGCACUGAUGU	3.69
mir-73	CACACACGACUGGACUUCCAUAUCGAGCCACAGCUAUCAACGAAUUUGCUGGCAAGAUGUAGGCAGUUCAGUUGU	2.77
mir-74	AAAUGGUUCA..(20 nt)..CUCUUUCCCAGCCUACAUCUCAACCUGGGCUGGCAAGAAAUGGCAGUCUACACGUUUUUCAACCA	6.46
mir-75	UUGCUUUGAAGAAUUGCAGUCGGUUGCAAGCUUAAAUACAAAUCCGAAUUGUUAUUAAAGCUACCAACCGGCUUCAAGUCUGAAAGAGCA	4.71
mir-76	UCCUGUCUGGGCUUCACAAUAGUCGAAUACCUUAAAUUUCAAAAUUUGGAUAUUCGUUGUUGAUGAAGCCUUGAUGGGGG	8.77
mir-77	GCCCGUUUGGAUGGUUGUGCUCUGAGGAAAUACGCACAGAAUGUCAUUUCAUCAGGCCAUAGCUGUCCAAAUUGGUAUAG	6.12
mir-78	AUAUUGUUUCAUAGUGUCCGUAAAAUAACUAGAUUUAUUUUGUAAAAACUAUUGGAGGCCUGGUUGUUUGUGCUG	0.27
mir-79	UCUCCGAUCUUUGGUGAUUCAGCUUCAAUGAUUGGCUACAGGUUUCUUUCAUAAAGCUAGGUUACCAAAGCUCGG	3.92
mir-80	UCGUUCGCUCAGCUUUCGACAUGAUUCUGAACAAUCCGCAAGCCCAUGUUGUUGAGAUCAUUAGUUGAAAGCCGAAUGAU	4.43
mir-81	GCCCAACAGUCGGUUUUCACCGUGAUCUGAGAGCAAUCCAAAAAUGCUUUUCUGAGAUCAUCGUGAAAGCUAGUUGUUGGUCUAC	6.48
mir-82	UUUAGCAACCGGUUUUCUCUGUGAUCUACAGAAUGACAGCUAAUCGUCUGAGAUCAUCGUGAAAGCCAGUUGUUU	5.13
mir-83	AACCACUGAAUUUAUGUGUGUACUUGACGGCCAACAAGAGCAUCGAUCUAGCACCAUAUAAAUUCAGUAAUUUCG	5.23
mir-84	tctcaacagaacagccgagttagttgaaacattgtggacattatagacagtcTACAATATTACATACTACCTCAg (antisense)	4.63
mir-85	GUCGGAGCCCGAUUUUUCAAUAGUUUGAAACCAGUGUACACAUAAAUGGUUACAAAGUAUUUGAAAAGUCGUGCUCUGAA	6.26
mir-86	gtgtcaaactccggcctaagcgaatctgagcccaggcttcatttcagaacatcgaaGACTGTGGCAAAGCATTCACTTAg (antisense)	3.82
mir-87	CAUCCGGCCGCCUGAUACUUUCGUCUCAACCUCGCUGUCAGAUUGGUCGUAGGUGAGCAAAGUUUCAGGUGUGCCGGAAC	6.81
mir-90	GCGCCAUUUCGAGCGGCUUUCAACGACGAUAUCAACCGACAACUCACACUUUUGCGUGUUGAUAUGUUGUUUGAAUGCCCCUUGAAUUGGAUGCC	6.04
mir-124	AUCUGGCAUGCACCCUAGUGACUUUAGUGGACAUCUAAGUCUUCCAACUAAGGCACGCGGUGAAUGCCACGUGGC	4.48
mir-228	CCUUAUCCCGUUCGCAAUGGCACUGCAUGAAUUCACGGCUAUGCAUAACGACAGACCGCGGAUCAUACGGUACCAUAGCGGACGGUGAUGAGGUU	5.14

Table 3: $SigZscr_e$ values were computed by scanning a set of fixed-length windows (55, 60, ..., 95, 100-nt) in steps of 2-nt along the sequence. The computed maximal $SigZscr_e$ was listed in the table.

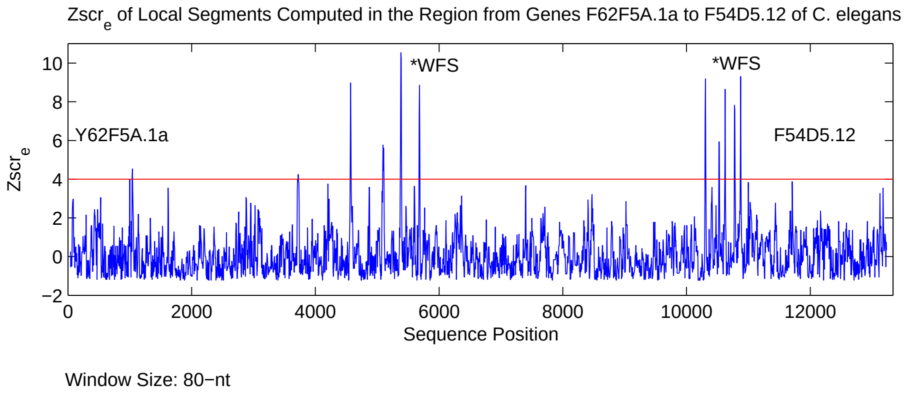

Figure 4: $Zscr_e$ of local segments computed in the region from genes F62F5A.1a to F54D5.12 of *C.elegans* chromosome II. The quantitative measure $Zscr_e$ was computed by moving a 80-nt window in steps of 5 nt from 5' to 3' along the sequence by EDscan. The plot was made by plotting $Zscr_e$ against the position of the middle base of the local overlapping segments. The end position (position 11527161 in the chromosome II sequence) of the gene Y62F5A.1a (antisense: 11533710-11527161) was numbered as position 1 in the plot. The five peaks clustered close to the gene F54D5.12 in the plot are coincident with the reported miRNAs, mir-35, mir-37, mir-38, mir-39, and mir-40. We also detected other WFSs that were clustered toward the gene Y62F5A.1a.

CONCLUSION

Rapid advances in computational biology and bioinformatics are providing new approaches to complex biological systems. Advances in systems biology and molecular medicine require combined efforts of bioinformaticists and molecular biologists. Such integrative approaches hold promise for elucidating gene function and RNA-based regulation of gene expressions. With the improvement of the integrating algorithms of statistical and computational tools of RNA folding, pattern search, sequence and structure comparison, computational methods can be used to discover FSRs that are associated with important biological properties. The need for these kinds of FSR discoveries is growing in proportion to the size of sequence databases, which are growing exponentially. The ncRNAs represent an important subset of the sequence databases in which potentially novel biological phenomena will be found. The existing tools, although already successful in finding interesting structural features of ncRNAs, can be improved further by future development.

ACKNOWLEDGMENTS

We thank J.H. Chen, B. Shapiro, J. Owens and K. Zhang for their cooperations in our projects. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

REFERENCES

1. Mattick, J.S. 2001, *EMBO Rep.,* **2,** 986.

2. Simons, R.W. and Grunberg-Manago M. eds. 1998, *RNA Structure and Function,* Cold Spring Harbor Lab. Press, New York.

3. Storz, G. 2002, *Science,* **296,** 1260.

4. Plasterk, R.H. 2002, *Science,* **296,** 1263.

5. Eddy, S.R. 2002, *Cell,* **109,** 137.

6. Gray, N.K. and Wickebs, M. 1998, *Annu. Rev. Cell Dev. Biol.,***14,** 399.

7. Bashirullah, A., Cooperstock, R.L. and Lipshitz, H.D. 1998, *Annu. Rev. Biochem.,* **67,** 335.

8. Cullen, B.R. 2003, *Trends in Biochemical Sciences,* **28,** 419.

9. Hellen, C.U.T. and Sarnow, P. 2001, *Genes & Development,***15,** 1593.

10. Macdonald, P.M. and Smibert, C.A. 1996, *Curr Opin Genet Dev.,* **6,**. *403.*

11. Hentze, M.W., Caughman, S.W., Casey, J.L., Koeller, D.M., Rouault, T.A. Harford, J.B. and Klausner, R.D. 1988, Gene, **72,** *201.*

12. Pasquinelli, A.E., Reinhart, B.J., Slack, F., Martindale, M.Q., Kuroda, M.I., Maller, B., Hayward, D.C., Ball, E.E., Degnan, B., Muller, P., Spring, J., Srinivasan, A., Fishman, M., Finnerty, J., Corbo, J., Levine, M., Leahy, P., Davidson, E. & Ruvkun, G. 2000, Nature, **408,** *86.*

13. Ambros, V. et al. 2003, RNA, **9,** *277.*

14. Gutell, R. R. 1993, Curr. Opin. Struct. Biol., **3,** *313.*

15. James, B.D., Olsen, G. & Pace, N.P. 1989, Methods in Enzymology, **180,** *227.*

16. Michel, F. and Westhof, E. 1990, J. Mol. Biol., **216,** *585.*

17. Nussinov, R. & Jacobson, A.B. 1980, Proc. Natl Acad. Sci. USA, **77,** *6309.*

18. Zuker, M. and Stiegler, P. 1981, Nucl. Acids Res. **9,** *133.*

19. Zuker, M. 1989, Science, **244,** *48.*

20. Mathews, D.H., Sabina, J., Zuker, M. and Turner, D.H. 1999, J. Mol. Biol., **288,** *911.*

21. Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, L., Tacker, M. and Schuster, P. 1994, Monatshefte Chem., **125,** *167.*

22. Sankoff, D. Kruskal, J.B., Mainville, S. and Cedergren, R.J. 1983, Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison, D. Sankoff, and J.B. Kruskal Ed) Addison-Wesley, 93.

23. Le, S.-Y., Chen J.-H. and Maizel Jr., J.V. 1993, Nucleic Acids Res., **21,** *2173.*

24. Laferriere, A., Gautheret, D. and Cedergren, R. 1994, Comput. Appl. biosci., **10,** *211.*

25. Billoud, B., Kontic, M. and Viari, A. 1996, Nucleic Acids Res., **24,** *1395.*

26. Grillo, G., Licciulli, F., Liuni, S., Sbisa, E. and Pesole, G. 2003, Nucleic Acids Res., **31,** *3608.*

27. Dsouza, M., Larsen, N. and Overbeek, R. 1997, Trends Genet., **13,** *497*

28. Macke, T.J., Ecker, D.J., Gutell, R.R., Gautheret, D., Case, D.A. and Sampath, R. 2001, Nucleic Acids Res., **29,** *4724.*

29. Gautheret, D. and Lambert, A. 2001, J. Mol Biol., **313,** *1003.*

30. Fichant,] G.A. and Burks, C. 1991, J. Mol Biol., **220,** *659.*

31. Pavesi, A., Conterio, F., Bolchi, A., Dieci, G. and Ottonello, S. 1994, Nucleic Acids Res., **22,** *1247.*

32. Lowe, T.M. and Eddy, S.R. 1997, Nucleic Acids Res., **25,** *955.*

33. Le, S.Y., Chen, J.H. and Maizel Jr., J.V. 1990, Structure & Methods: Human Genome Initiative and DNA Recombination., R.H. Sarma and M.H. Sarma (Ed), Adenine Press, Schenectady, 127.

34. Le, S.Y., Chen, J.H., Konings, D. and Maizel Jr., J.V. 2003, Bioinformatics, **19,** *354.*

35. Le, S.Y., Chen, J.H. and Maizel Jr., J.V. 2003, Proceedings of the 2003 IEEE Bioinformatics Conference, CSB2003 Stanford, California, 190.

36. Lim, L.P., Lau, N.C., Weinstein, E.G., Abdelhakim, A., Yekta, S., Rhoades, M.W., Burge, C.B. and Bartel, D.P. 2003, Genes & Development, **17,** *991.*

37. Lai, E.C., Tomancak, P., Williams, R.W. and Rubin, G.M. 2003, Genome Biology, **4,** *R42.*

38. Moore, P.B. 1999, Annu. Rev. Biochem., **68,** *287.*

39. Hermann, T. and Patel, D.J. 2000, Structure, **8,** *R47.*

40. Leontis, N.B. and Westhof, E. 1998, J. Mol. Biol., **283,** *571.*

41. Gautheret, D., Konings, D. and Gutell, R.R. 1994, J. Mol. Biol., **242,** *1.*

42. Draper, D.E. 1996, Trends Biochem Sci., **21,** *145.*

43. Schultes, E.A., Hraber, P.T. and LaBean, T.H. 1999, J Mol. Evol., **49,** *76.*

44. Collins, G., Le, S.Y. and Zhang, K. 2001, Information Sciences, **139,** *59.*

45. Le, S.Y., Zhang, K. and Maizel Jr., J.V. 2002, Nucl. Acids Res., **30,** *3574.*

46. Malim, M.H., Hauber, J., Le, S.Y., Maizel Jr., J.V. and Cullen, B.R. 1989, Nature, **338,** *254.*

47. Le, S.Y., Chen, J.H. and Maizel Jr., J.V. 1989, Nucl. Acids Res., **17,** *6143.*

48. Hanly, S.M., Rimsky, L.T., Malim, M.H., Kim, J.H., Hauber, J., Dodon, M., Le, S,Y., Maizel Jr., J.V., Cullen, B.R. and Greene, W.C. 1989, Genes & Development, **3,** *1534.*

49. Malim, M.H., Bohnlein, S., Fenrick, R., Le, S.Y., Maizel Jr., J.V. and Cullen,B.R. 1989, Proc. Natl. Acad. Sci. USA, **86,** *8222.*

50. Dayton, E.T., Konings, D.A., Powell, D.M., Shapiro, B.A., Butini, L., Maizel Jr., J.V. and Dayton, A.I. 1992, J. Virol., **66,** *1139.*

51. Yang, J., Bogerd, H., Le, S.Y. and Cullen, B.R. 2000, RNA, **6,** *1551.*

52. Gorodkin, J., Heyer, I.J. and Stormo, G.D. 1997, Proc. Int. Conf. Intell. Syst. Mol. Biol., **5,** *120.*

53. Juan, V. and Wilson, C. 1999, J. Mol. Biol, **289,** *935.*

54. Hofacker, I.L. and Stadler, P.F. 1999, Comp. & Chem, **23,** *401.*

55. Ruan, J., Stormo, G.D. and Zhang, W. 2004, Bioinformatics **20,** *58.*

56. Le, S.Y., Zhang, K. nd Maizel Jr., J.V. 1995, Comp. Biomed. Res., **28,** *53.*

57. Chen, J.H., Le, S.Y. nd Maizel Jr., J.V. 2000, Nucl. Acide Res., **28,** *991.*

58. Le, S.Y., Maizel Jr., J.V. and Zhang, K. 2004, Proceedings of the 2004 IEEE Bioinformatics Conference, CSB2004 Stanford, California, in press.

59. Le, S.Y. and Maizel Jr., J.V. 1997, Nucl. Acids Res., **25,** *362.*

60. Akiri, G., Elroy-Stein, O., Nahari, D., Finkelstein, Y., Le, S.Y. and Levi, B.Z. 1998, Oncogene, **17,** *227.*

61 Sella, O., Gerlitz, G. Le, S.Y. and Elroy-Stein, O. 1999, Mol. Cell Biol., **19,** *5429.*

62. Pesole, G., Liuni, S., Grillo, G., Ippedico, M., Larizza, A., Makalowski, W. and Saccone, C. 1999, Nucl. Acids Res., **27,** *188.*

63. Carter, R.J., Dubchak, I and Holbrook, S.R. 2001, Nucl. Acids Res., **29,** *3928.*

64. Szymanski, M., Barciszewska, M.Z., Erdmann, V.A. and Barciszewski, J. 2003, Biochem. J., **371,** *641.*